Coursera Course7 Final Project

Erica Pehrsson

12/19/2018

Executive summary

Improving the efficiency of motor vehicles (as measured by miles travelled per gallon of gasoline, mpg) has important benefits both for the environment and the consumer. Here, we use the 1974 mtcars dataset of motor vehicles to identify factors that explain the mpg measure, focusing in particular on whether a car has automatic or manual transmission. Without controlling for other factors, cars with manual transmission are significantly more efficient than those with automatic transmission (Wilcox test, difference in mean mpg 7.245). However, when controlling for other factors such as vehicle weight and number of cylinders using multivariate linear regression, mpg is not a significant predictor. Therefore, transmission type does not have a significant impact on mpg.

Exploratory data analyses

The dataset used in this analysis is mtcars, which contains 11 pieces of data for 32 cars from the 1974 edition of ‘’Motor Trend’’ US magazine.

Overview of variables

V vs. straight engine (vs) and automatic vs. manual transmission (am) are binary values, while number of cylinders (cyl; 4/6/8), number of forward gears (gear; 3/4/5), and number of carburetors (carb; 1/2/3/4/6/8) are restricted to 3-6 values each. The remaining variables, displacement (disp), gross horsepower (hp), rear axle ratio (drat), weight (wt), and 1/4 mile time (qsec), are continuous.

```
apply(mtcars,2,function(x) length(unique(x)))
summary(mtcars)
```

Miles per gallon distribution

Across all 32 models, miles per gallon (mpg) ranges from 10.4 to 33.9, with a median of 19.2 and a roughly normal distribution.

Manual transmission vehicles have a significantly higher mpg than those with automatic transmissions (Wilcox test, $p < 0.005$; difference in means, 7.25). (Appendix Plot 1)

```
summary(mtcars$mpg)
ddply(mtcars,.(am),function(x) summary(x$mpg))
wilcox.test(mpg~am,data=mtcars)
```

Model selection

The first model, comparing only mpg with automatic/manual transmission, has a coefficient of determination (R2) of 0.3598, suggesting that ~36% of the variation in mpg is explained by the model. Transmission is a significant variable. As expected, manual transmissions increase the mpg by an average of 7.245 compared to automatic transmissions.

```
m1 = lm(mpg~as.factor(am),data=mtcars)
summary(m1)
```

Next, I included variables that are highly correlated (absolute value) with mpg. (Appendix Plot 2) The most highly correlated variables with mpg are weight (wt), number of cylinders (cyl), and displacement (disp).

```
cors = melt(as.matrix(cor(mtcars)))
a = cors[which(cors$Var1 == "mpg"),]
a[order(abs(a$value)),]
```

Controlling for weight in the model removes the significance of transmission as a variable (automatics tend to be heavier than manual cars). (Appendix Plot 3) For every half ton of additional weight, the mpg decreases by 5.35, and the variable is highly signficant. ANOVA indicates that the improvement to the model with the addition of weight is significant.

```
m2 = lm(mpg~as.factor(am)+wt,data=mtcars)
summary(m2)
anova(m1,m2)

wilcox.test(wt~am,data=mtcars)
```

Including the number of cylinders also significantly improves the model, and weight and number of cylinders are both significant predictors.

```
m3 = lm(mpg~as.factor(am)+wt+cyl,data=mtcars)
summary(m3)
anova(m2,m3)
```

However, displacement is not significant when correcting for the other variables, likely because it is highly correlated with both weight and number of cylinders.

```
m4 = lm(mpg~as.factor(am)+wt+cyl+disp,data=mtcars)
summary(m4)
anova(m3,m4)

cors[which(cors$Var1 == "disp"),]
```

Therefore, I chose model 3 (mpg versus transmission, weight, and number of cylinders) for my model. (Appendix Plot 4) The R2 for the model is 0.83, suggesting that the model explains much of the variation in mpg.

Residual plots and diagnostics

I used residual plots to determine whether there are any biases in the model. (Appendix Plot 5) There is a slight negative correlation between the fitted values and the residuals, with the exception of a few outliers (Toyota Corolla and Fiat 128), suggesting that there may be some residual variation unexplained by the model.

The residuals also appear normally distributed, with the exception of the outliers mentioned before, and are homoscedastic. However, none of the outliers has a Cook's distance of greater than 0.5, suggesting that none should be excluded from the model.

```
sort(cooks.distance(m3))
```

It is worth noting, however, that the Chrysler Imperial has a higher dfbeta for the weight variable than other vehicles. This is likely because it has an usually high mpg for its weight compared to other vehicles with automatic transmission. (Appendix Plot 6)

```
dfbetas(m3)
```

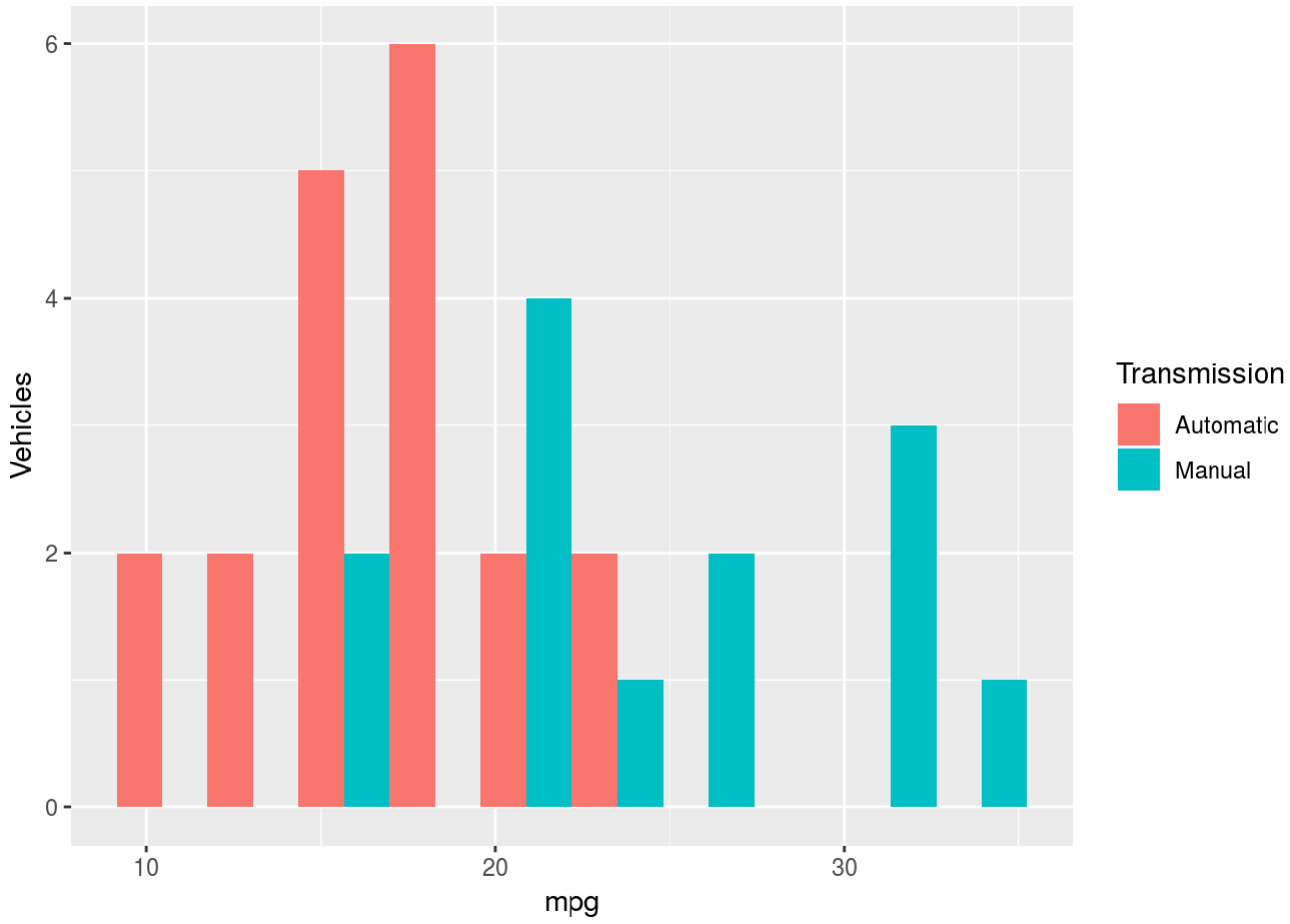
Uncertainty

Although both weight (wt) and number of cylinders (cyl) are significant variables in the model, there is a high degree of uncertainty for both. Although on average an increase in weight of 1kb results in a 3.1251 decrease in mpg, and an increase in the number of cylinders results in a 1.5102 decrease in mpg, the 95% confidence intervals for these values are broad.

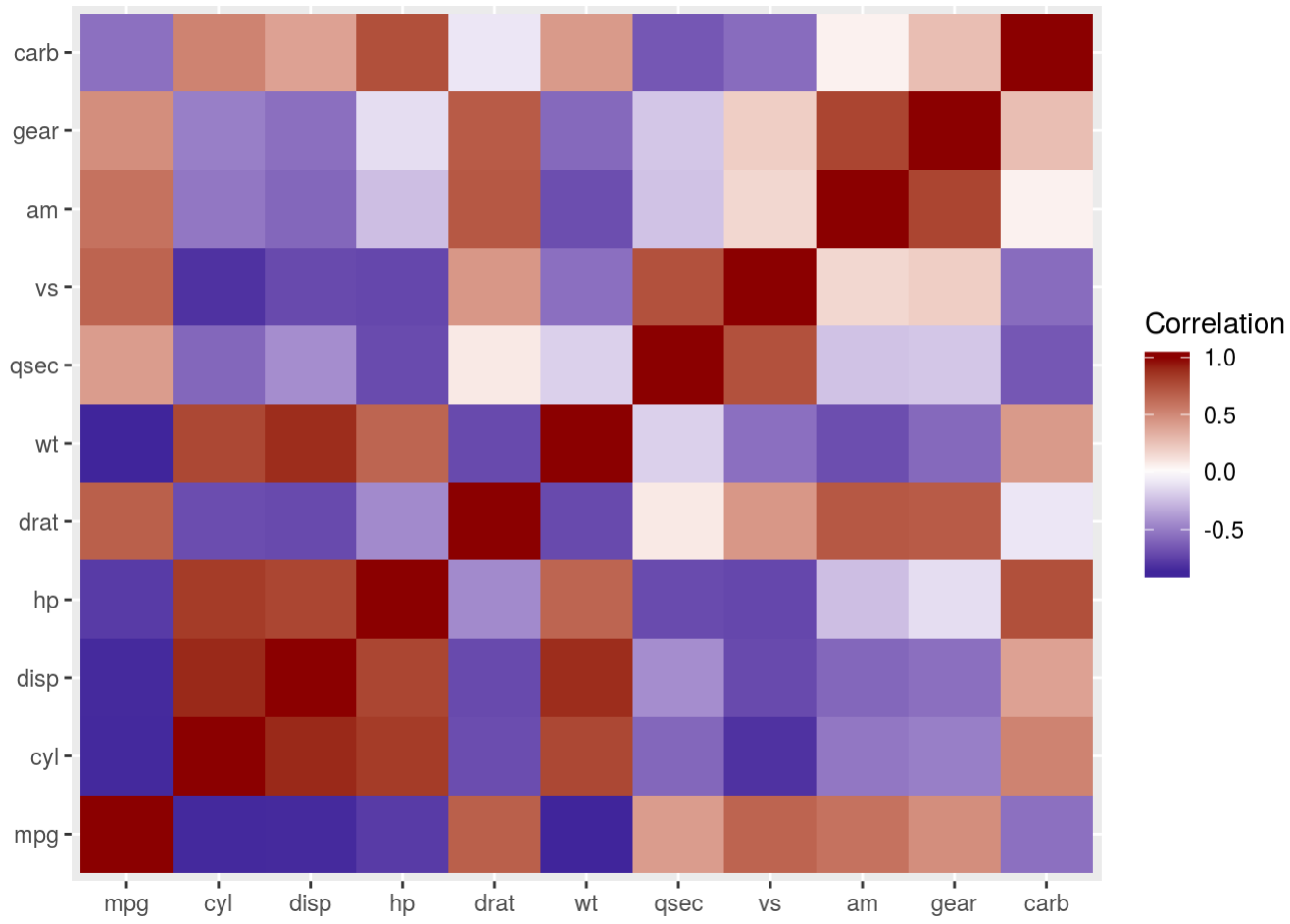
```
confint(m3)
```

Appendix

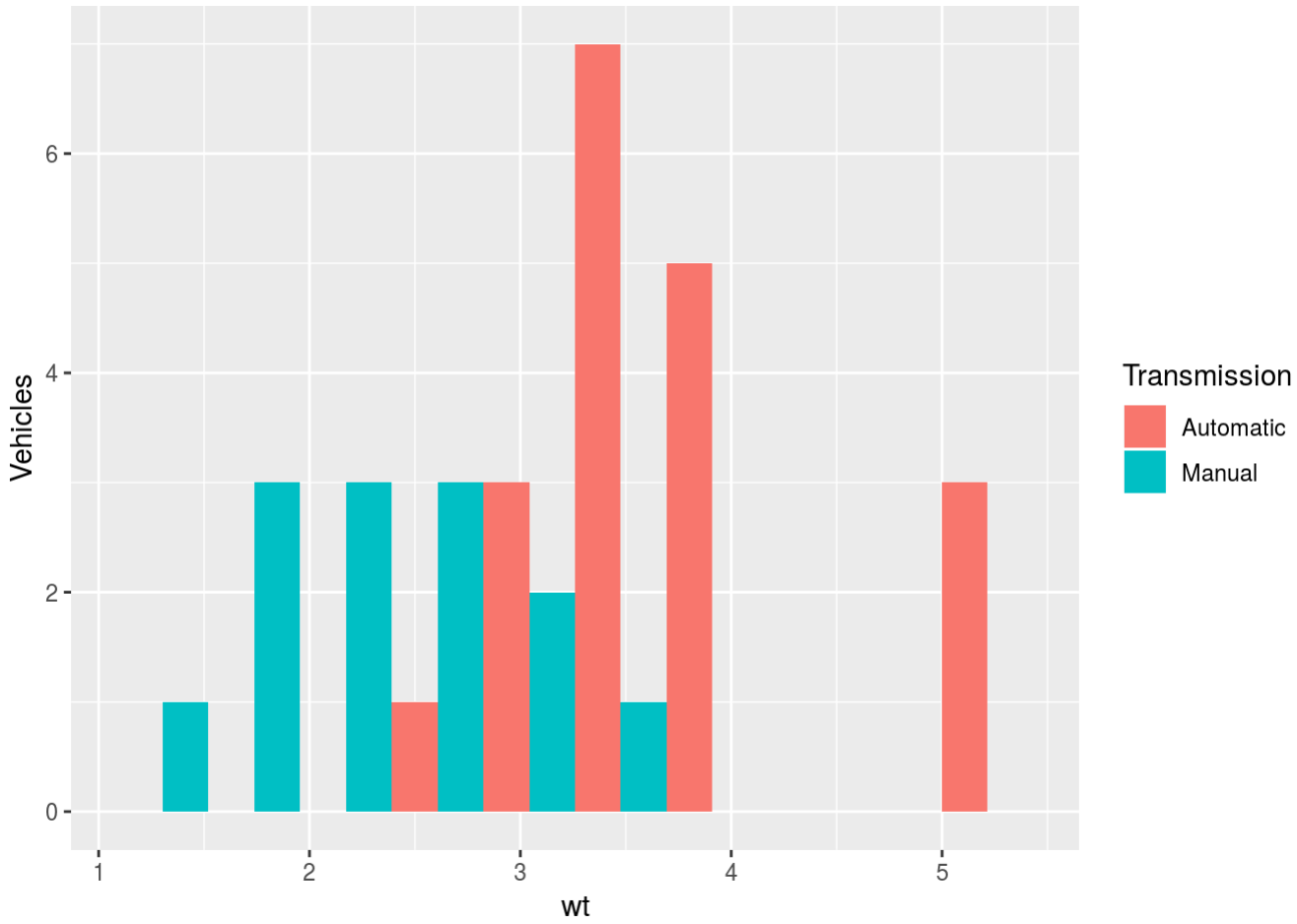
Distribution of mpg by transmission type



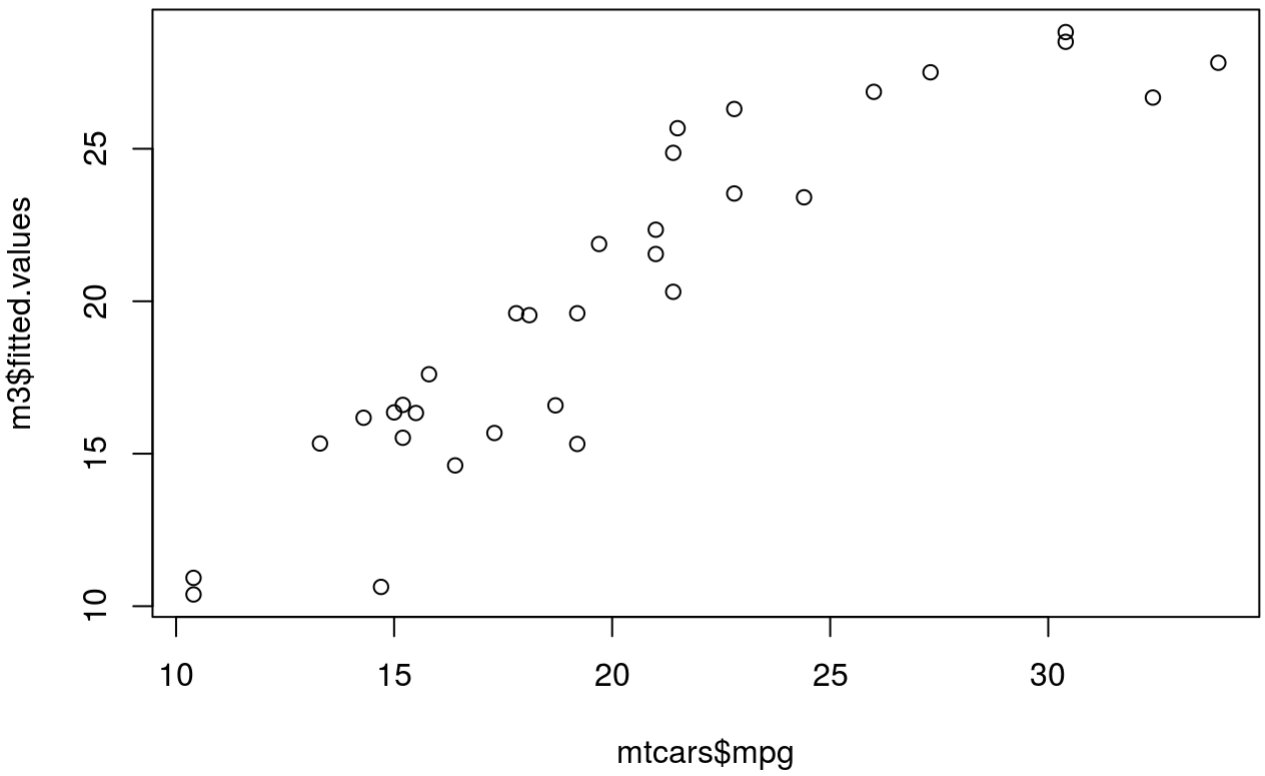
Variable correlation (Pearson)



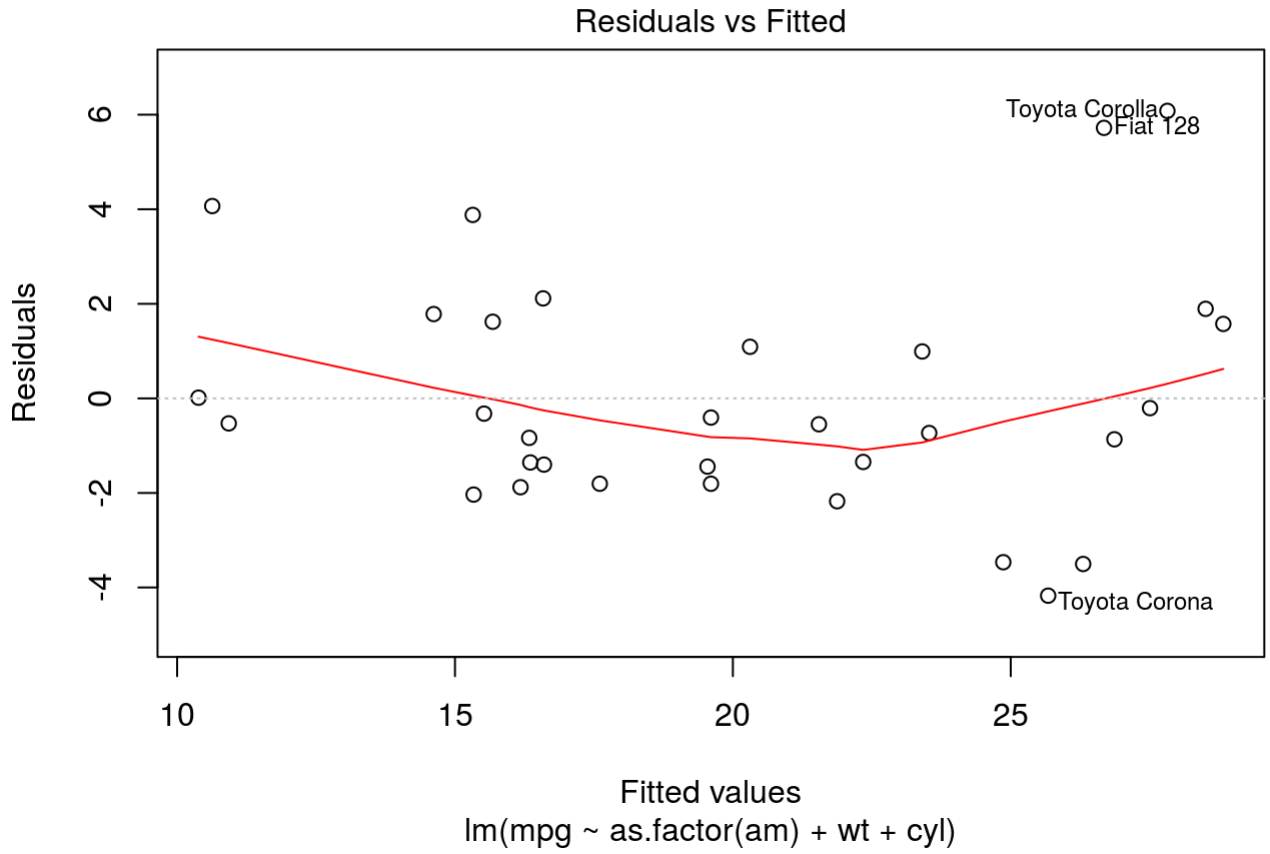
Distribution of weight by transmission type

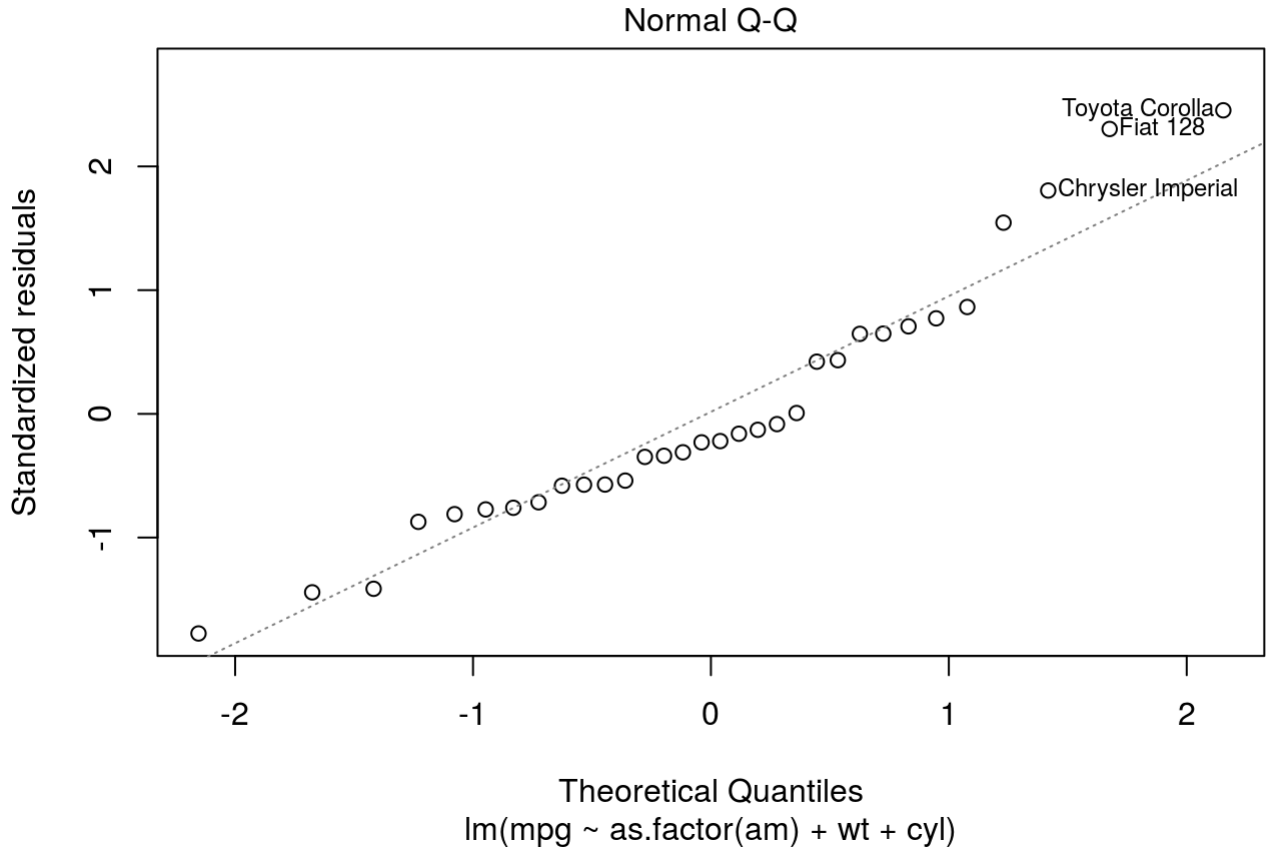


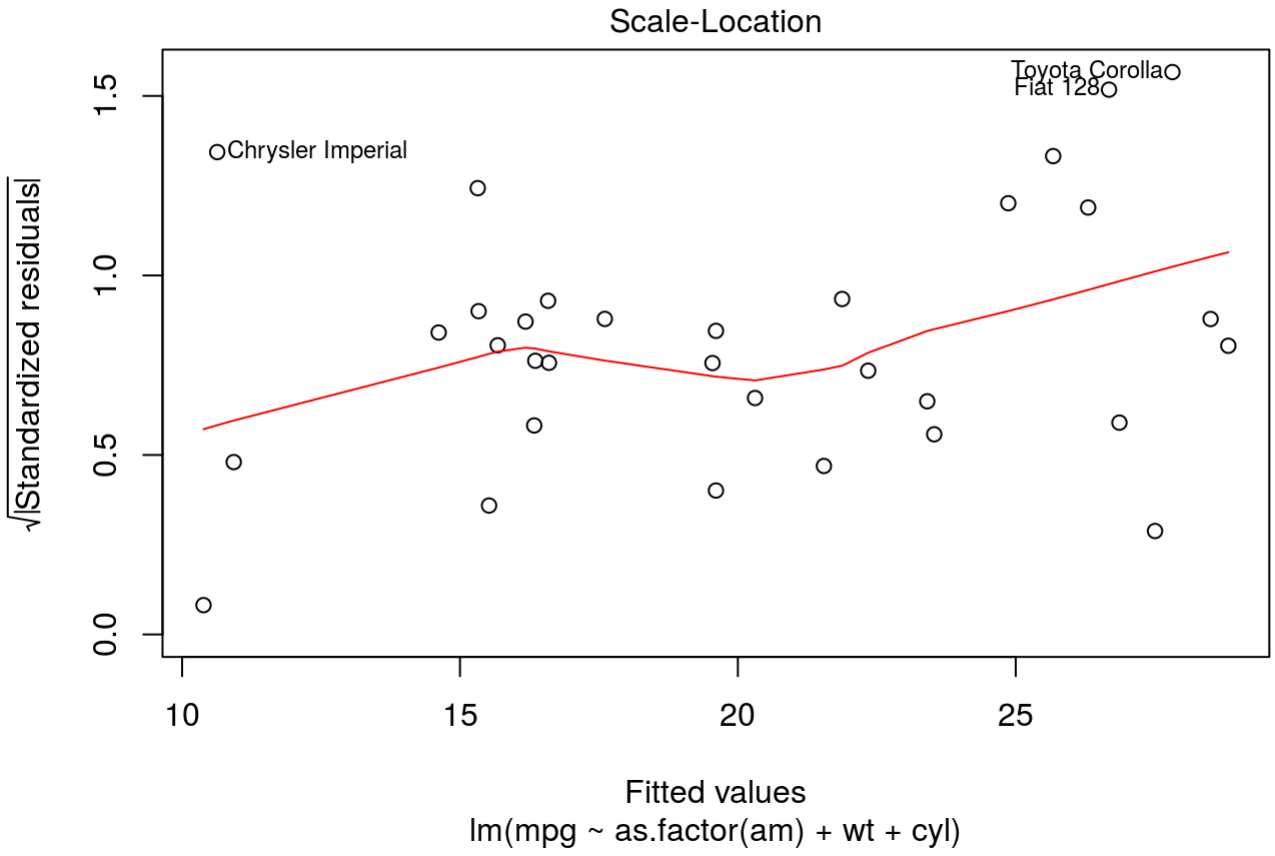
Real vs. fitted values for model 3

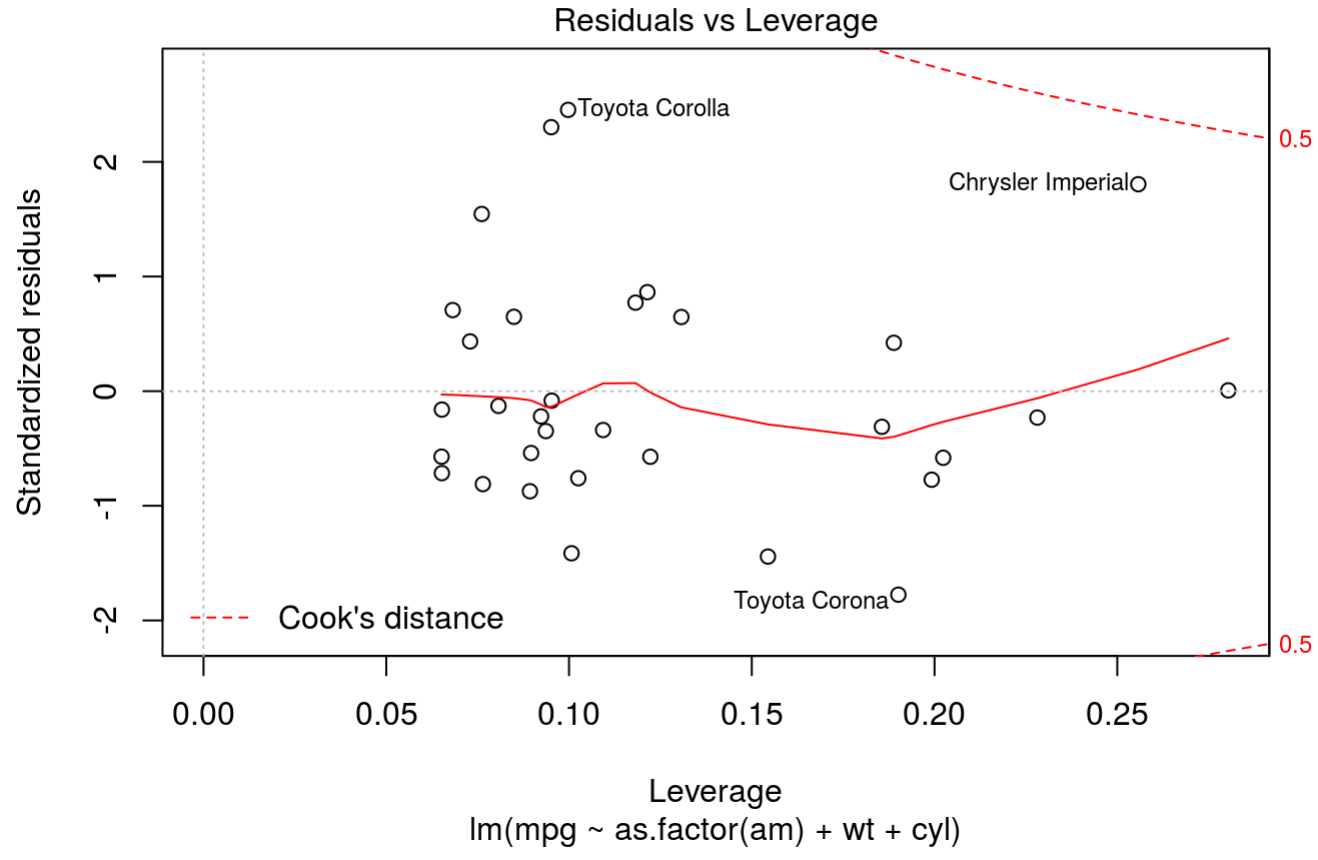


Residual diagnostic plots for model 3

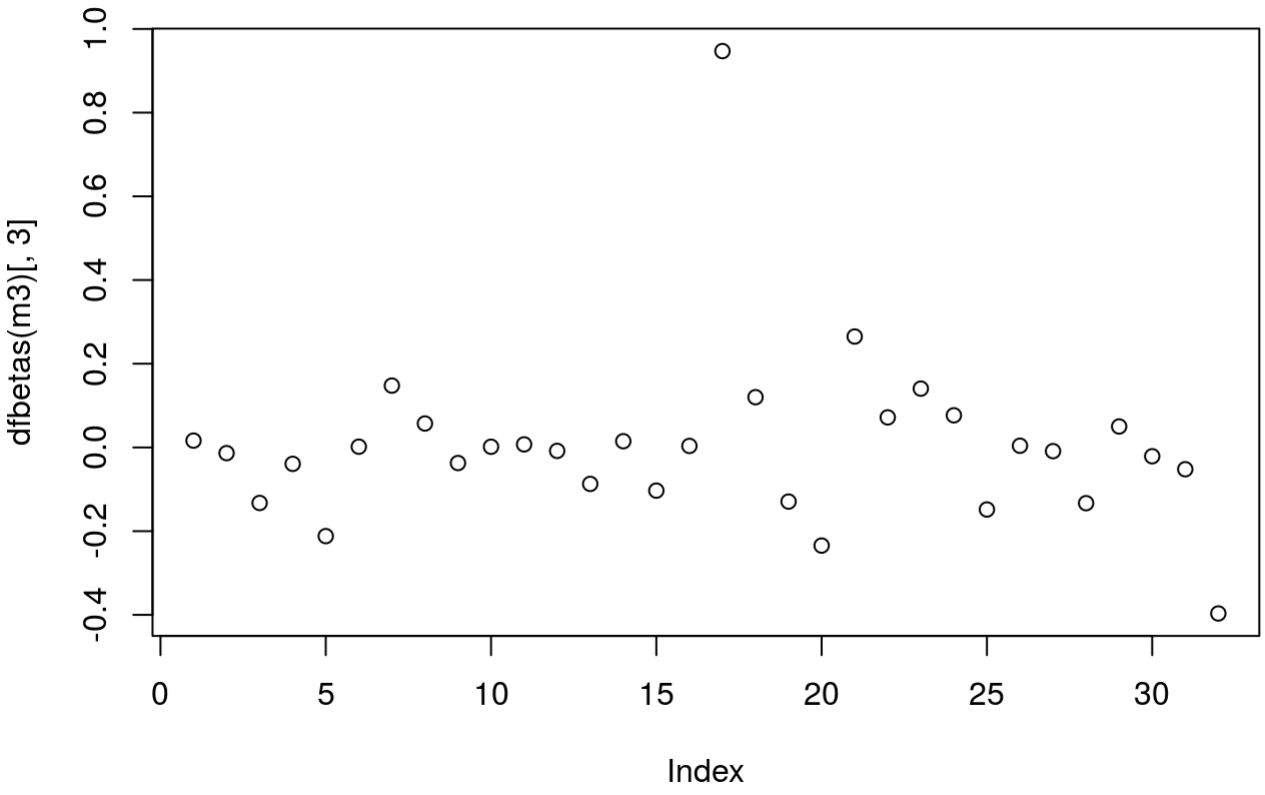

Residuals vs Fitted
Toyota Corolla
Fiat 128
Toyota Corona
Residuals
6
4
2
0
-2
-4
10
15
20
25
Fitted values
lm(mpg ~ as.factor(am) + wt + cyl)

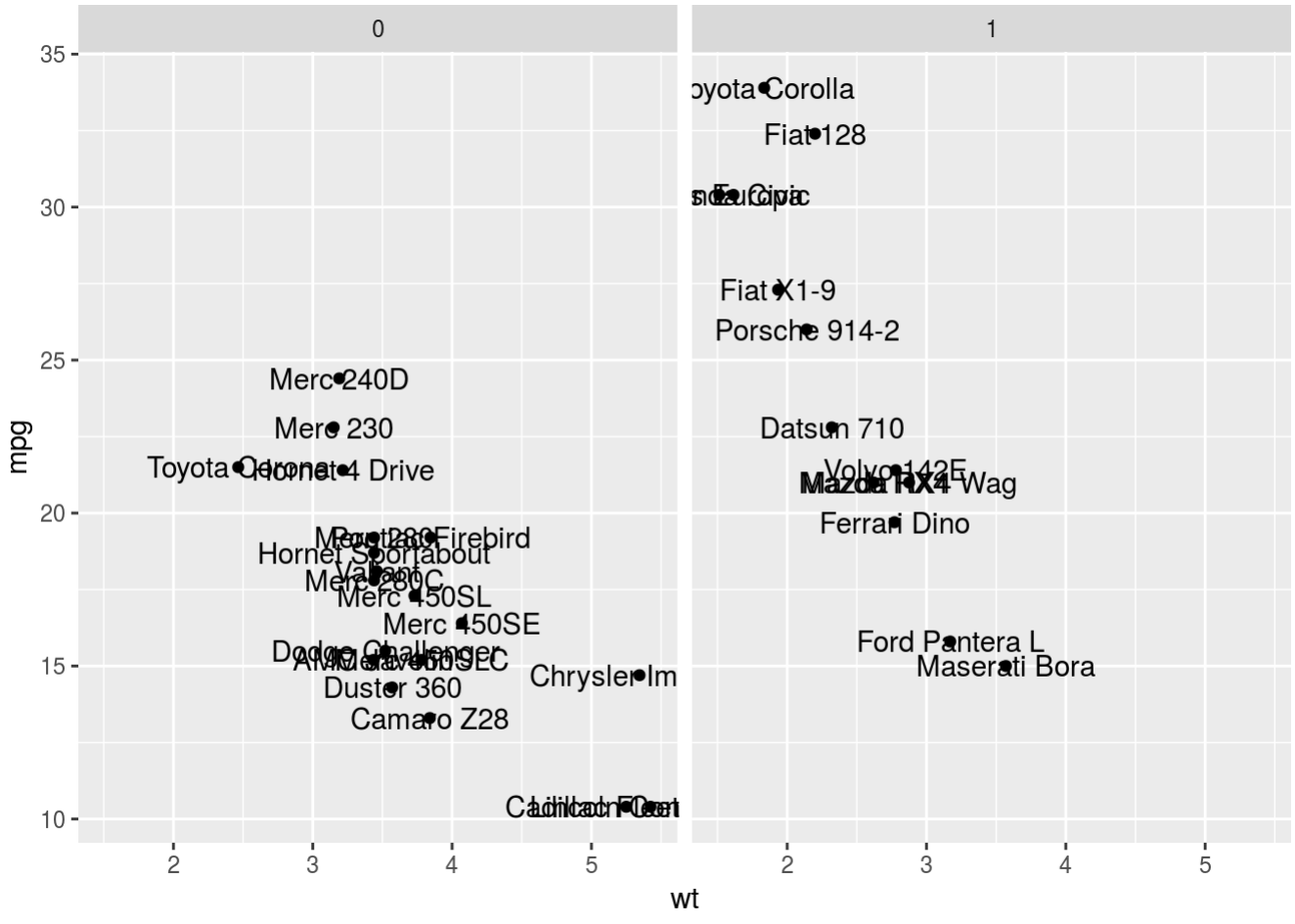

Normal Q-Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Standardized residuals
2
1
0
-1
-2
-1
0
1
2
Theoretical Quantiles
lm(mpg ~ as.factor(am) + wt + cyl)


Scale-Location
Toyota Corolla
Fiat 128
Chrysler Imperial
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
Fitted values
lm(mpg ~ as.factor(am) + wt + cyl)


Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5
0.5
Standardized residuals
-2
-1
0
1
2
0.00
0.05
0.10
0.15
0.20
0.25
Leverage
lm(mpg ~ as.factor(am) + wt + cyl)

Dfbetas for vehicle weight


dfbetas(m3)[, 3]
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
5
10
15
20
25
30
Index


0
1
mpg
35
30
25
20
15
10
2
3
4
5
wt
Merc 240D
Merc 230
Hornet 4 Drive
Merc 280C
Merc 450SL
Merc 450SE
Duster 360
Camaro Z28
Chrysler Im
Fiat 128
Fiat X1-9
Porsche 914-2
Datsun 710
Ferrari Dino
Ford Pantera L
Maserati Bora